

Biologia Molecular Computacional
IBI5035/QBQ2507 - 2023

Bancos de Dados Biológicos

Eduardo Moraes Rego Reis
Instituto de Química - USP

Estudos ômicos geram grande quantidade de dados moleculares

DNA → RNA → Proteína

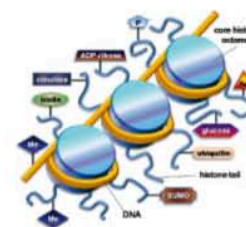
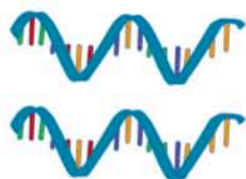
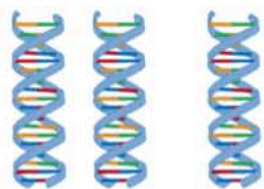
modificação química no
DNA/RNA/histonas

Genômica

Transcritômica

Proteômica

Epigenômica



> Replicase polyprotein 1ab, P19811
MATFSATGFG GSFVRDWSLD LPDACE-
HGAG LCCEVDGSTL CAECFRGCEG M

• Sequenciamento de DNA

• Microarranjos
• RNA-Seq

• Espectrometria de massas

• ChIP-Seq
• Sequenciamento DNA bissulfito

• ChIP-Seq

Sequencia de genes e genomas

• Sequencia de RNAs
• Abundância de RNAs

• Sequencia, abundância, modificações pós-tradução de proteínas

• Modificações de histonas, metilação de DNA

• Sítios de ligação de fatores de transcrição
• Estados de ativação da cromatina

Tecnologias de células únicas

Contents [hide]

- 1 [Meta databases](#)
- 2 [Model organism databases](#)
- 3 [Nucleic acid databases](#)
 - 3.1 [DNA databases](#)
 - 3.2 [Gene expression databases \(mostly microarray data\)](#)
 - 3.3 [Phenotype databases](#)
 - 3.4 [RNA databases](#)
- 4 [Amino acid / protein databases](#)
 - 4.1 [Protein sequence databases](#)
 - 4.2 [Protein structure databases](#)
 - 4.3 [Protein model databases](#)
 - 4.4 [Protein-protein and other molecular interactions](#)
- 5 [Signal transduction pathway databases](#)
- 6 [Metabolic pathway and protein function databases](#)
- 7 [Additional databases](#)
 - 7.1 [Exosomal databases](#)
 - 7.2 [Mathematical model databases](#)
 - 7.3 [Taxonomic databases](#)
 - 7.4 [Radiologic databases](#)
- 8 [Wiki-style databases](#)
- 9 [Specialized databases](#)
- 10 [References](#)
- 11 [External links](#)

Categorias de bancos de dados

Primários

Dados gerados experimentalmente
ex. sequencias/estruturas de DNA, RNA, proteínas.

Dados genomicos brutos e normalizados de RNAseq “bulk” e célula única

Função de arquivamento.

Secundários

dados resultantes da análise, interpretação e curagem de dados primários

ex. famílias de proteínas, genes alvos de microRNAs, variação/regulação/função de sequencias de DNA (SNPs, SNVs, Indels)

Agregação de informação

Search NCBI

TP53



Search

Results found in 36 databases for **TP53**

Literature

Bookshelf	956	Books and reports
MeSH	17	Ontology used for PubMed indexing
NLM Catalog	17	Books, journals and more in the NLM Collections
PubMed	16,219	Scientific and medical abstracts/citations
PubMed Central	36,735	Full-text journal articles
PubMed Health	95	Clinical effectiveness, disease and drug reports

Genes

EST	1,202	Expressed sequence tag sequences
Gene	4,379	Collected information about gene loci
GEO DataSets	6,080	Functional genomics studies
GEO Profiles	193,438	Gene expression and molecular abundance profiles
HomoloGene	13	Homologous gene sets for selected organisms
PopSet	35	Sequence sets from phylogenetic and population studies
UniGene	141	Clusters of expressed transcripts

Genetics

ClinVar	1,722	Human variations of clinical significance
dbGaP	0	Genotype/phenotype interaction studies
dbVar	3,660	Genome structural variation studies
GTR	545	Genetic testing registry
MedGen	48	Medical genetics literature and links
OMIM	360	Online mendelian inheritance in man
SNP	6,148	Short genetic variations

Proteins

Conserved Domains	4	Conserved protein domains
Identical Protein Groups	254	Protein sequences grouped by identity
Protein	3,988	Protein sequences
Protein Clusters	3	Sequence similarity-based protein clusters
Sparcle	18	Functional categorization of proteins by domain architecture
Structure	171	Experimentally-determined biomolecular structures

Genomes

Assembly	0	Genome assembly information
BioCollections	0	Museum, herbaria, and other biorepository collections
BioProject	451	Biological projects providing data to NCBI
BioSample	827	Descriptions of biological source materials
Clone	532	Genomic and cDNA clones
Genome	29	Genome sequencing projects by organism
GSS	6	Genome survey sequences
Nucleotide	8,621	DNA and RNA sequences
Probe	1,533	Sequence-based probes and primers
SRA	2,430	High-throughput sequence reads

Chemicals

BioSystems	3,874	Molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	447	Bioactivity screening studies
PubChem Compound	0	Chemical information with structures, information and links
PubChem Substance	621	Deposited substance and chemical information



Cancer Single-cell Expression Map

A public database dedicating to collecting, analyzing, visualizing single-cell RNA-Seq data of human cancers. Multi-level analyses were performed to deeply explore the tumor microenvironment of different types of human cancers and a comprehensive online analyze platform was equipped in the database.

- Home
- Project Browse
- Search
- Analyze
- Documentation
- References
- Downloads
- Help

Data release 1.0

Quick Search



(e.g.Lung Adenocarcinoma, TP53, ENSG00000186891)

Keyword Cloud



28 PROJECTS



20 CANCER TYPES



638,341 CELLS

GEPIA 2

- Home
- FUNCTIONS**
 - Expression Analysis
 - Custom Data Analysis
- EXTRAS**
 - Docs
 - Examples
 - Dataset Sources
 - Deconvolution Analysis



- Single Gene Analysis**
- Cancer Type Analysis
- Custom Data Analysis
- Multiple Gene Analysis

Enter gene/isoform name:

The indicators in search box are "symbol" or "alias (newest symbol)".

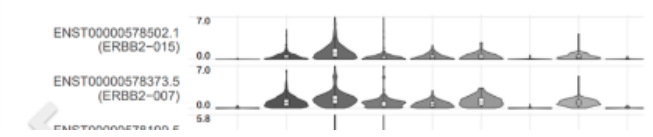
GoPIA!

- Profile
- Boxplots
- Stage Plots
- Survival Analysis
- Similar



Isoform structure

Isoform usage profiling



Bancos de dados para anotação, visualização e
integração de dados genômicos com
informações da literatura

Ex. UCSC Genome Browser, ENSEMBL

Tipos de dados representados em visualizadores de dados genômicos

Estrutura e padrão de expressão gênica

- isoformas de splicing
- estrutura exon/intron
- regiões codificadoras/UTRs
- expressão em diferentes tecidos/células

Regiões regulatórias no DNA

- Promotores, enhancers
- sítios de ligação de fatores de transcrição, histonas

Variação na sequência do DNA/RNA

- SNPs, indels,
- mutações, edição de RNA

Conservação evolutiva

alinhamento de DNA entre espécies

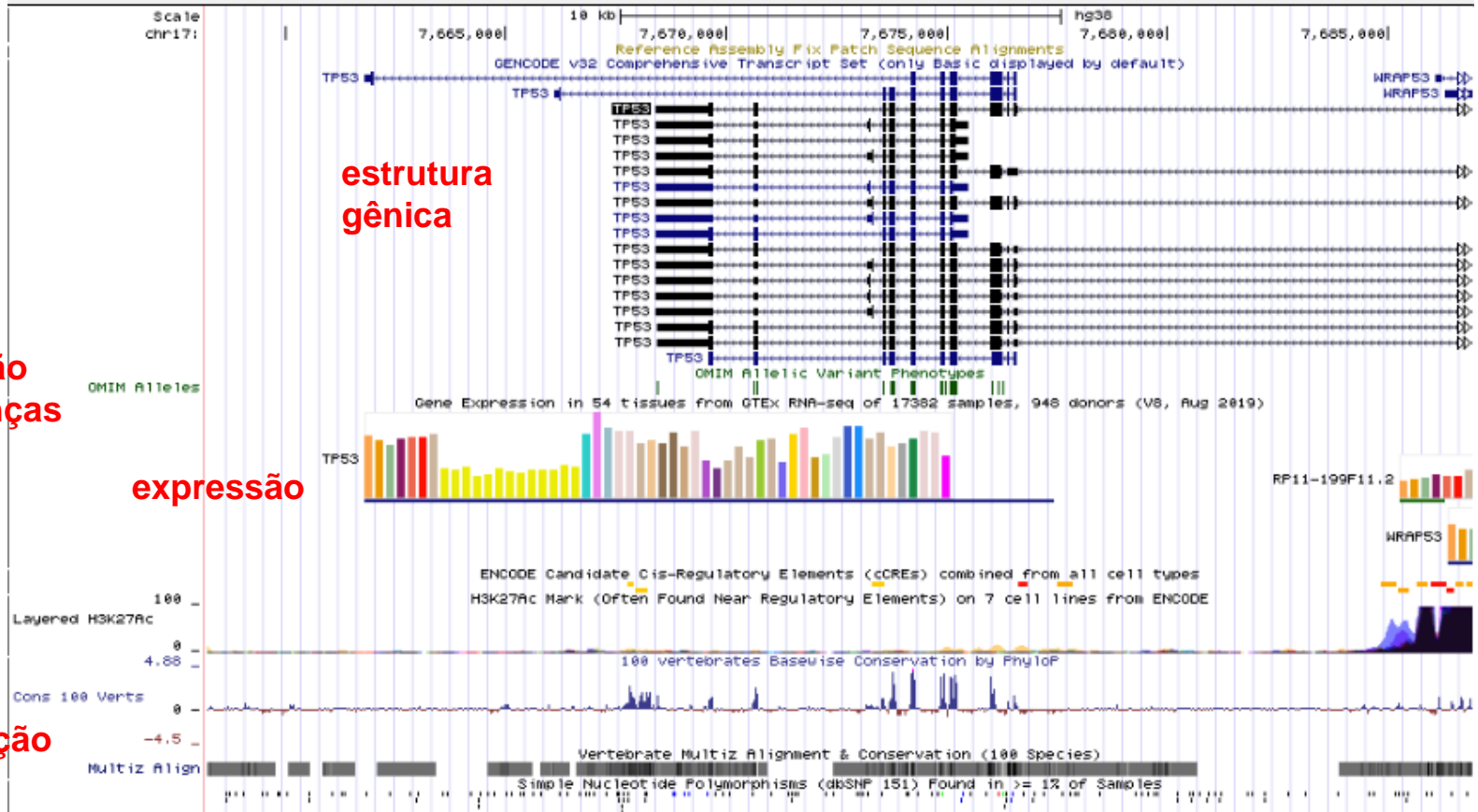
<http://genome.ucsc.edu/cgi-bin/hgGateway>

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr17:7,658,228-7,686,933 28,706 bp. enter position, gene symbol, HGVS or search terms go

chr17 (p13.1) 13.3 13.2 p13.1 17p12 17p11.2 17q11.2 17q12 21.31 17q22 23.2 24.3q25.1 17q25.3



estrutura
gênica

associação
com doenças

expressão

regulação

conservação

variação



Informação está organizada em tracks de anotação

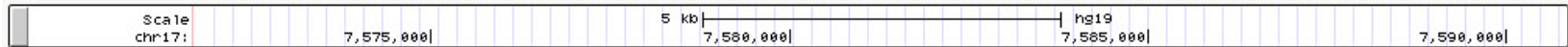
- chromosome band
- gap locations
- known genes
- predicted genes
- phenotype and disease
- enhancer/promoter data
- microarray/expression data
- evolutionary conservation
- SNPs and structural variation
- repeated regions
- more...

Como controlar a visibilidade das tracks

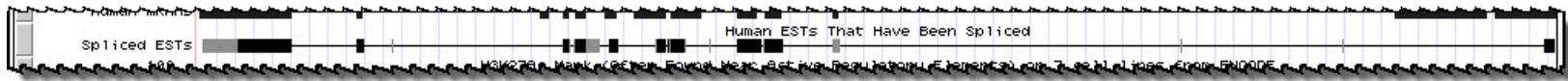
- Hide: removes a track from view

Spliced ESTs

dense ▾
hide
dense
squish
pack
full



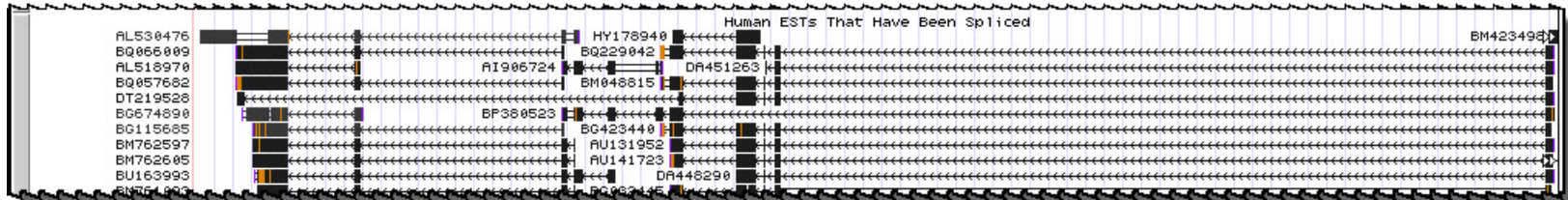
- Dense: all items collapsed into a single line



- Squish: each item = separate line, but 50% height + packed



- Pack: each item separate, but efficiently stacked (full height)



- Full: each item on separate line (may need to zoom to fit)



Os tracks estão conectados com diferentes camadas de informação

Scale chr17: 7,575,000 | 5 kb | hg19 | 7,590,000

UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

TP53

Click the item

New description web page opens

Gene TP53 (uc002gij.3) Description and Page Index

Description: Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.

Summary (NM_001276760): This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in this gene are associated with a variety of human cancers, including hereditary cancers such as Li-Fraumeni syndrome. Alternative splicing of this gene and the use of alternate promoters result in multiple transcript variants and isoforms. Additional isoforms have also been shown to result from the use of alternate translation initiation codons (PMIDs: 12032546, 20937277). [provided by RefSeq, Feb 2013].

Transcript (Including UTRs)
Position: chr17:7,571,720-7,590,868 Size: 19,149 Total Exon Count: 11 Strand: -

Coding Region
Position: chr17:7,572,927-7,579,569 Size: 6,643 Coding Exon Count: 8

Page Index	Sequence and Links	UniProtKB Comments	Genetic Associations	CTD
Microarray	RNA Structure	Protein Structure	Other Species	GO Annotations
Pathways	Other Names	GeneReviews	Model Information	Methods

Data last updated: 2013-06-14

Sequence and Links to Tools and Databases

Genomic Sequence (chr17:7,571,720-7,590,868)	mRNA (may differ from genome)	Protein (354 aa)			
Gene Sorter	Genome Browser	Protein FASTA	Table Schema	BioGPS	CGAP

Example: click your mouse anywhere on the TP53 line

Formatos para representação de dados genômicos

<http://genome.ucsc.edu/FAQ/FAQformat.html#format1>

General formats

- **Axt format**
- **BAM format**
- **BED format**
- **BED detail format**
- **bedGraph format**
- **barChart and bigBarChart format**
- **bigBed format**
- **bigGenePred table format**
- **bigPsl table format**
- **bigMaf table format**
- **bigChain table format**
- **bigNarrowPeak table format**
- **bigWig format**
- **Chain format**
- **CRAM format**
- **GenePred table format**
- **GFF format**
- **GTF format**
- **HAL format**
- **interact and bigInteract format**
- **MAF format**
- **Microarray format**
- **Net format**
- **Personal Genome SNP format**
- **PSL format**
- **VCF format**
- **WIG format**

- **Desafio: Dados de diferentes naturezas: estrutura gênica, regulação, expressão, variação, etc.**

Em comum:

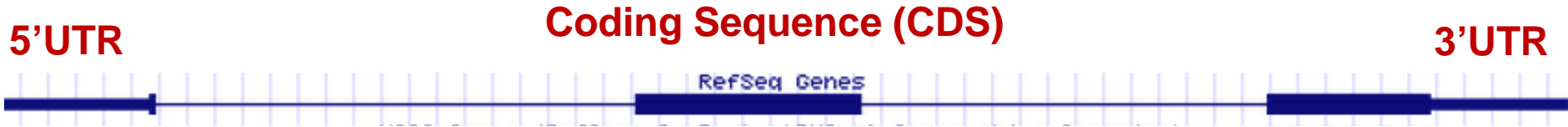
- **contem linhas de colunas separadas por tabulação**
- **baseados em coordenadas genômica**

Formato BED (Browser Extensible Data)

BED 3, BED 6, BED 12

chr	chromStart	chromEnd	name	score	strand
chr1	67051159	67163158	NM_024763	0	-
chr1	67075869	67163158	NM_207014	0	-
chr1	50286272	50440127	NM_001144777	0	+
chr1	41744622	42156965	NM_024503	0	-
chr1	8335050	8800286	NM_001042681	0	-
chr1	33544953	33559286	NM_001080438	0	-
chr1	8335050	8406334	NM_001042682	0	-
chr1	8335050	8800286	NM_012102	0	-
chr1	41744622	42156965	NM_001127714	0	-
chr1	58718978	58785034	NM_145243	0	-
chr1	16761508	16812687	NM_017940	0	-
chr1	100590610	100723514	NM_033313	0	+
chr1	100590610	100737609	NM_033312	0	+
chr1	100590610	100758421	NM_003672	0	+
chr1	75444662	75849387	NM_152697	0	-
chr1	117404471	117447014	NM_003594	0	+
chr1	75440403	75849387	NM_001130058	0	-
chr1	92268120	92301681	NM_173567	0	+
chr1	184532027	184550311	NM_001127709	0	+
chr1	184532027	184550311	NM_001127710	0	+
chr1	184532027	184550311	NM_005807	0	+
chr1	184532027	184550311	NM_001127708	0	+
chr1	184547408	184611080	NM_003292	0	-
chr1	167747815	167822393	NM_000130	0	-
chr1	243199793	243317771	NM_032328	0	+
chr1	243200253	243317771	NR_026588	0	+
chr1	243199906	243355153	NR_026586	0	+
chr1	243200253	243355153	NR_026587	0	+
chr1	226462453	226615574	NM_052843	0	+
chr1	226462453	226633198	NM_001098623	0	+
chr1	243200253	243355153	NM_001143943	0	+
chr1	226462453	226633198	NM_001271223	0	+
chr1	6246918	6341591	NM_181866	0	-
chr1	6767970	7752353	NM_015215	0	+
chr1	2026014	2106694	NM_001033582	0	+

Arquivo BED com coordenadas do gene MYC



chr8 128817496 128822862 NM_002467 0 + 128818021 128822386 0 3 555,772,1039, 0,2179,4327,

campo 1: chr8	cromossomo
campo 2: 128817496	coordenada início do gene
campo 3: 128822862	coordenada fim do gene
campo 4: NM_002467	nome do gene (GeneID)
campo 5: 0	score (definição da representação gráfica)
campo 6: +	fita em que o gene se localiza
campo 7: 128818021	coordenada início da CDS
campo 8: 128822386	coordenada fim da CDS
campo 9: 0	itemRGB (definição da cor no gráfico)
campo 10: 3	número de exons
campo 11: 555,772,1039,	tamanho dos exons (em nt)
campo 12: 0,2179,4327,	coordenada de início dos exons (em relação ao campo 2)

<http://genome.ucsc.edu/cgi-bin/hgGateway>

Através do Genome Browser é possível:

- Fazer consultas textuais utilizando o nome, símbolo, localização genômica
- Alinhar sequências de nucleotídeo/proteína no genoma (BLAT)
- Cruzar informações de “tracks” de anotação e selecionar regiões de DNA de interesse (“Table Browser”)

consultas textuais utilizando o
nome, símbolo, localização
genômica

Gateway: Start Page for a Basic Search

UCSC Genome Browser Gateway

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Browse/Select Species

POPULAR SPECIES

Human Mouse Rat Fruitfly Worm Yeast

Enter species or common name

REPRESENTED SPECIES

Human Chimp Bonobo Gorilla Orangutan Gibbon Crab-eating macaque Rhesus Baboon (anubis) Baboon (hamadryas) Marmoset Squirrel monkey Tarsier Mouse lemur Bushbaby Mouse Rat Chinese hamster Kangaroo rat Squirrel Naked mole-rat Guinea pig

Find Position

Human Assembly
Dec. 2013 (GRCh38/hg38)

Position/Search Term
Enter position, gene symbol or search terms
Current position: chr9:133,252,000-133,280,861

GO

Human Genome Browser - hg38 assembly view sequences

UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: GRCh38 Genome Reference Consortium Human Reference 38 (GCA_000001405.15)
Assembly date: Dec. 2013
GenBank accession ID: GCA_000001305.2
NCBI Genome information: NCBI genome/51 (Homo sapiens)
NCBI Assembly information: NCBI assembly/883148 (GRCh38/GCA_000001405.15)
BioProject information: NCBI Bioproject: 31257

Search the assembly:

- By position or search term: Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- By gene name: Type a gene name into the "search term" box, choose your gene, and then press "submit" to go directly to the assembly location associated with that gene. More information.
- By track type: Click the "track search" button to find Genome Browser tracks that match your search criteria. More information.

Download sequence and annotation data:

- Using rsync (recommended)
- Using FTP
- Using HTTP
- Data use conditions and restrictions
- Acknowledgments

Assembly Details

The GRCh38 assembly is the first major revision of the human genome released in more than four years. As with the previous GRCh37 assembly, the

■ Use this Gateway to search:

- Gene names, symbols, IDs
- Chromosome number: chr7, or region: chr11:1038475-1075482
- Keywords: kinase, receptor

Sample Search for Human TP53

- Sample search: human, February 2009 assembly, tp53

UCSC Genes

[TP53 \(uc002gij.3\) at chr17:7571720-7590868](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.

[TP53 \(uc031qvp.1\) at chr17:7571720-7579937](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 2, mRNA.

[TP53 \(uc010vug.3\) at chr17:7577851-7590868](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 3, mRNA.

uc002gij.3

[TP53 \(uc002gin.3\) at chr17:7577499-7590868](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 4, mRNA.

[TP53 \(uc002gin.3\) at chr17:7571720-7590868](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 5, mRNA.

[TP53 \(uc002gii.2\) at chr17:7571720-7578811](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 6, mRNA.

[TP53 \(uc002gih.3\) at chr17:7569404-7579937](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 7, mRNA.

[TP53 \(uc002gig.1\) at chr17:7565097-7579937](#) - Homo sapiens tumor protein p53 (TP53), transcript variant 8, mRNA.

[TP53TG30 \(uc021tgv.1\) at chr16:32264650-32267243](#) - Homo sapiens TP53 target 30 (TP53TG30), transcript variant 1, mRNA.

[TP53TG38 \(uc010caz.3\) at chr16:33262120-33264719](#) - Homo sapiens TP53 target 38 (TP53TG38), transcript variant 3, mRNA.

[TP53TG38 \(uc010caz.3\) at chr16:33205585-33208179](#) - Homo sapiens TP53 target 38 (TP53TG38), transcript variant 4, mRNA.

[TP53TG1 \(uc003uip.4\) at chr7:86954664-86974831](#) - Homo sapiens TP53 target 1 (non-coding) (TP53TG1), transcript variant 4, mRNA.

[TP53TG5 \(uc002zmv.4\) at chr20:44002528-44007033](#) - Homo sapiens TP53 target 5 (TP53TG5), transcript variant 4, mRNA.

RefSeq Genes

[TP53 at chr17:75](#)

[TP53 at chr17:75](#)

[TP53 at chr17:75](#)

[TP53 at chr17:75](#)

[TP53 at chr17:75](#)

[TP53 at chr17:75](#)

[TP53 at chr17:75](#)

[TP53AIP1 at chr17](#)

[TP53AIP1 at chr17](#)

Comprehensive Gene Annotation Set from ENCODE/GENCODE

[TP53 at chr17:7565097](#)

[TP53 at chr17:7569404](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

[TP53 at chr17:7571720](#)

Ensembl Genes

[TP53INP1 at chr8:95938200-95961594](#) - (ENST00000448464)

[TP53TG1 at chr7:86974513-86974802](#) - (ENST00000542586)

[TP53TG1 at chr7:86954666-86974802](#) - (ENST00000421293)

[TP53TG1 at chr7:86954541-86974831](#) - (ENST00000359941)

[TP53RK at chr20:45313004-45318418](#) - (ENST00000372114)

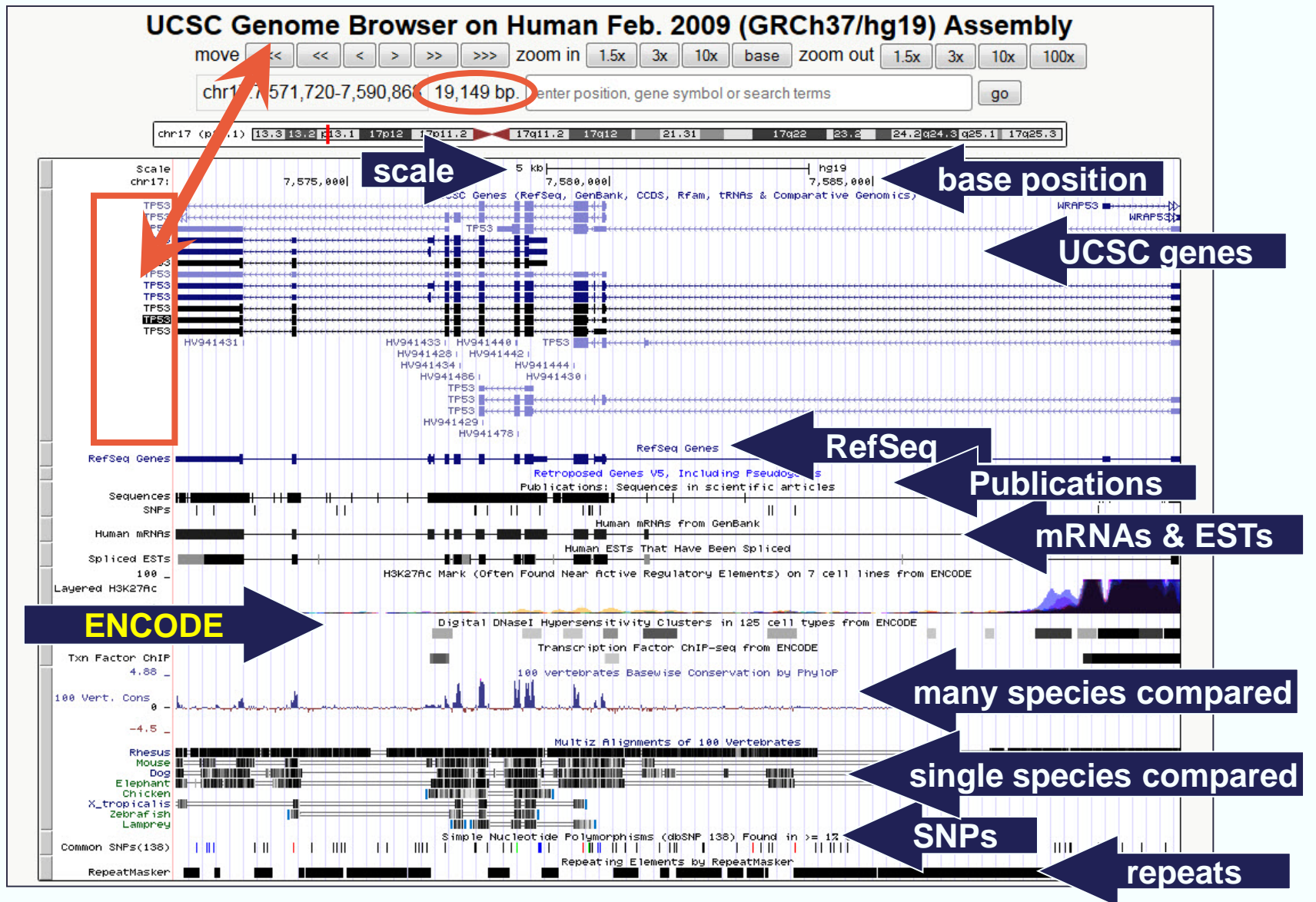
[TP53TG5 at chr20:44005837-44034634](#) - (ENST00000488588)

[TP53TG5 at chr20:44002528-44007033](#) - (ENST00000537995)

[TP53TNR2 at chr20:33282588-33287147](#) - (ENST00000414082)

- Select from results list; or goes to a viewer page, if unique

Sample Genome Viewer Image, TP53 Region



Visual Cues on the Genome Browser



Tick marks; a single location (STS, SNP)

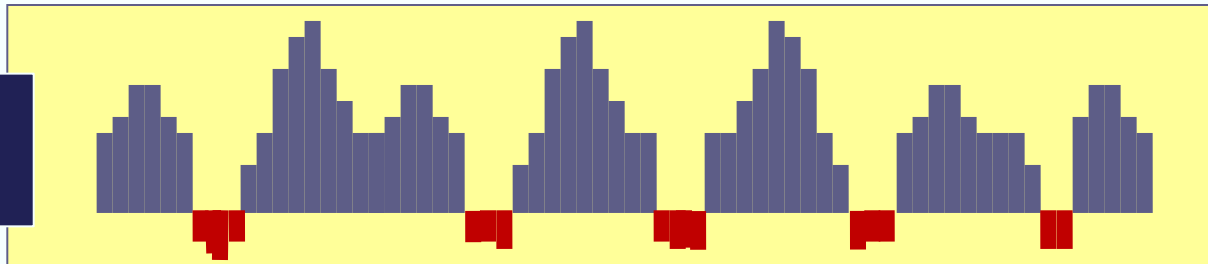


Intron and direction of transcription <<< or >>>

Track colors may have meaning—for example, UCSC Gene track:

- If there is a corresponding PDB entry = black
- If there is a corresponding reviewed/validated seq = dark blue
- If there is a non-RefSeq seq = lightest blue

Vert.
cons.



Wiggle

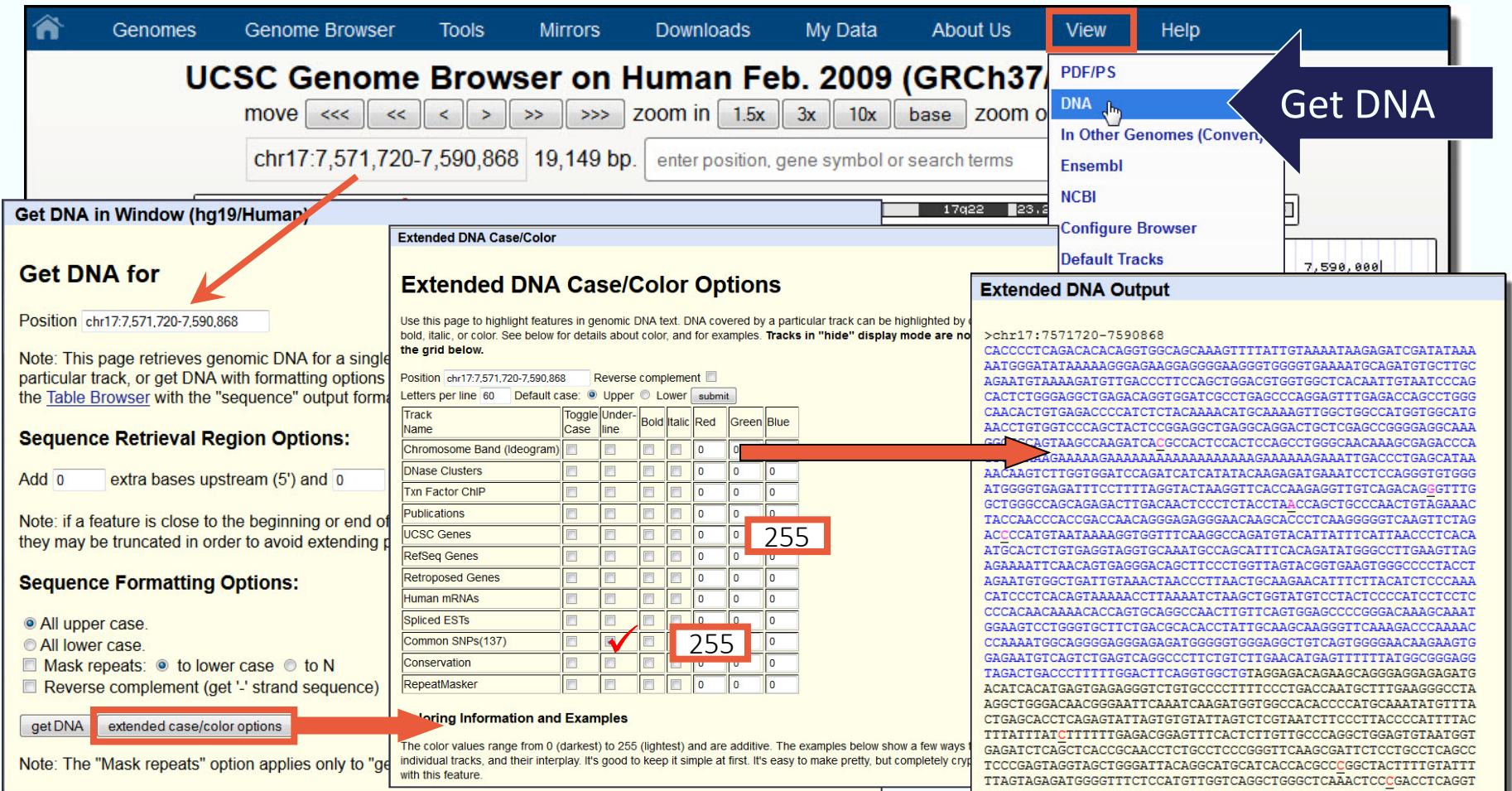
height of a blue bar is increased likelihood of conservation,
red indicates a likelihood of faster-evolving regions

Alignment indications (Conservation pairs: “chain” or “net” style)

- Alignments = boxes, Gaps = lines



Get DNA, with Extended Case/Color Options



UCSC Genome Browser on Human Feb. 2009 (GRCh37)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out

chr17:7,571,720-7,590,868 19,149 bp. enter position, gene symbol or search terms

View Help

- DNA
- In Other Genomes (Convert)
- Ensembl
- NCBI
- Configure Browser
- Default Tracks

Get DNA in Window (hg19/Human)

Get DNA for

Position chr17:7,571,720-7,590,868

Note: This page retrieves genomic DNA for a single particular track, or get DNA with formatting options the [Table Browser](#) with the "sequence" output format.

Sequence Retrieval Region Options:

Add 0 extra bases upstream (5') and 0

Note: if a feature is close to the beginning or end of they may be truncated in order to avoid extending p

Sequence Formatting Options:

- All upper case.
- All lower case.
- Mask repeats: to lower case to N
- Reverse complement (get '-' strand sequence)

Extended DNA Case/Color

Use this page to highlight features in genomic DNA text. DNA covered by a particular track can be highlighted by bold, italic, or color. See below for details about color, and for examples. Tracks in "hide" display mode are not the grid below.

Position chr17:7,571,720-7,590,868 Reverse complement

Letters per line: 60 Default case: Upper Lower

Track Name	Toggle Case	Underline	Bold	Italic	Red	Green	Blue
Chromosome Band (Ideogram)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
DNase Clusters	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Txn Factor ChIP	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Publications	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
UCSC Genes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	255
RefSeq Genes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Retroposed Genes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Human mRNAs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Spliced ESTs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Common SNPs(137)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	255
Conservation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
RepeatMasker	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0

Extended DNA Output

```
>chr17:7571720-7590868
CACCCCTCAGACACACAGGTGGCAGCAAAAGTTTATTGTAATAAATAGAGATCGATATAAA
AATGGGATATAAAAAGGGAGAGAGGGGAAGGGTGGGTGAAATGCAAGATGTGCTTGC
AGAAATGTAAGAGATGTTGACCCCTCCAGCTGGACGCTGGTGGCTCAAAATGTAATCCCG
CACTCTGGGAGGCTGAGACAGGTGGATCGCCTGAGCCCAAGGATTTGAGACCAGCTGGG
CAACACTGTGAGACCCCTCTCTACAAAACATGAAAAGTTGGCTGGCCATGGTGGCATG
AACCTGTGGTCCAGCTACTCCGGAGGCTGAGGCAGGACTGCTCGAGCCGGGGAGCCAAA
AGCCAGATGAAAGCCAGATCAAGCCACTCCACTCCAGCCTGGGCACAAAAGGAGACCA
CCAGAGTAAAGAAAAAAGAAAAAAGAAAAAAGAAAAAAGAAAAAAGAAATGACCCCTGAGCATAA
AAGAGTCTTGGTGGATCCAGATCATCATATAACAGAGATGAAATCCTCCAGGGTGGGG
ATGGGTTGAGATTTCTTTTAGGTAAGTACTAGGTTACCACAGAGTTGTGACAGAGGTTTG
GCTGGGCCAGCAGACTTGACAACCTCCCTCTACTCAACAGCTGCCCACTGTAGAAAC
TACCACCCACCCGACCAACAGGAGAGGGGAACAACGACCCCTCAAGGGGTCAGTTCTAG
ACCCATGTAATAAAGAGTGGTTTCAAGCCAGATGTACATATTTTCAATTAACCCCTCACA
ATGCACCTGTGAGGTTAGGTGCAAAATCCAGCAATTCACAGATATGGCCCTTGAAGTTAG
AGAAAATTCACAGTGAAGGACAGCTTCCCTGGTGTAGTCGGTGAAGTGGGCCCTCAACT
AGAAATGGGCTGATTTGAAACTAACCCCTAACCTGCAAGAACATTTCTCATCTCCCCAA
CATCCCTCACAGTAAAAACCTTAAATCTAAGCTGGTATGTCTACTCCCCATCCTCCTC
CCACAAACAAACACCAGTGCAGGCCAATCTGTCAGTGGAGCCCGGGACAAGGCAAAAT
GGAAGTCTGGTGGTCTTCTGACGCACACCTATTGCAAGCAAGGTTCAAGAGCAACAAAC
CCAAAATGGCAGGGGAGGAGAGATGGGGGTGGGAGGCTGTCAAGTGGGGAACAAAGAGT
GAGAATGTCAGTCTGAGTCAGGCCCTCTGTCTGTGAACATGAGTTTATGTCGGGGAGG
TAGACTGACCCCTTTTGGACTTCAGTGGCTGTAGGAGACAGAAGCAGGAGGAGAGATG
ACATCACATGAGTGAAGGGTCTGTGCCCTTTCCCTGACCAATGCTTTGAAGGSCCTA
AGGCTGGACAAAGGGAAATCAAATCAAGATGGTGGCCACCCCATGCAAAATATGTTTA
CTGAGCACCTCAGAGTATTAGTGTGATTAGCTCTGTAATCTTCCCTTACCCCATTTTAC
TTTATTTATCTTTTTTGAAGCGAGTTTCACTCTGTTGCCAGGCTGGAGTGAATGGT
GAGATCTCAGCTCAGCCAACTCTGCCTCCGGGTTCAAGCGATTCTCCTGCTCAGCC
TCCCGAGTAGTGTAGTGGGATACAGGCATGATCACCAGCCCGGCTACTTTTGTATTT
TTAGTAGAGATGGGTTTCTCCATGTTGGTCAGGCTGGCTCAAACTCCGACCTCAGGT
GATCCACTCGCCTTGGCTCCAGAGTGTGGGATTCGTGAGCCACTGGCCCGGGCCCTT
TACCCATTTTATATAAAGAACTGAGTTTACGCGGGGTCACTAGGACCTGCGCGGTG
CATGGCAGGGCTGAGTATGACCTGAAACTCTGGCTGATTCAATACAAATTAAT
AGGCCCTCCTTGAACCCCTCCAGCTCTGGCTGGGATGCGGAGAAATGGCAAGAAATG
ATCCACACTCGTCCCTGGGTTGGATGTTCTGTGGATACACTGAGGCAAGAAATGGTTA
TAGGATTCACCGGAGAAAGACTAAAAAATGTCTGTGAGGCTGGGACCCCAATGAGAT
GGGTCAGCTGCCTTTGACCATGAAGGCAGGATGAGAAATGGAATCCTATGGCTTTCCAC
CTAGGAAGGCAGGGGAGTAGGGCCAGGAAGGGGCTGAGGTCACCTCACCTGGAGTGGCC
TGCTCCCCCTGGCTCCTTCCAGCCTGGGCATCCTTGAGTTCCAAGGCCTATTCAAGCT
CTCGAAACATCTCGAAGGCTCAGCCGCCAGGATCTGCAGCAACAGAGGGGAGGAGAG
TAAGTATATACAGTACCTGAGTTAAAAGATGGTTCAAGTTACAATGTTTGAATTTAT
```

- Use the View → DNA link at the top
- Plain or Extended options
- Change colors, fonts, underline, etc.

Alinhar sequencias de
nucleotídeo/proteína no genoma
(BLAT)

Accessing the BLAT Tool

The screenshot shows the UCSC Genome Bioinformatics website. The main navigation bar includes links for Genomes, Blat, Tables, Gene Sorter, PCR, VisiGene, Session, FAQ, and Help. The left sidebar contains links for Genome Browser, Ebola, Blat, Table Browser, Gene Sorter, In Silico PCR, and Genome Graphs. The main content area is titled "About the UCSC Genome Bioinformatics Site" and contains a welcome message, a list of tools, and contact information. A right-hand sidebar features a "RESOURCES" section with a link to "BLAT—The BLAST-Like Alignment Tool" by W. James Kent, along with a logo for "GENOME RESEARCH".

UCSC Genome Bioinformatics

Genomes - **Blat** - Tables - Gene Sorter - PCR - VisiGene - Session - **FAQ** - Help

Genome Browser
Ebola
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference for a large collection of genomes. It also provides portals to [ENCODE](#) data project.

We encourage you to explore these sequences with our tools. The [Genome Browser](#) showing the work of annotators worldwide. The [Gene Sorter](#) shows expressed genes that can be related in many ways. [Blat](#) quickly maps your sequence to convenient access to the underlying database. [VisiGene](#) lets you browse the images to examine expression patterns. [Genome Graphs](#) allows you to update

The UCSC Genome Browser is developed and maintained by the Genome Browser within the [UC Santa Cruz Genomics Institute](#) and the Center for Biomolecular University of California Santa Cruz ([UCSC](#)). If you have feedback or questions, feel free to contact us on our [public mailing list](#).

Vol. 12, Issue 4, 656-664, April 2002

RESOURCES
BLAT—The BLAST-Like Alignment Tool
W. James Kent
Department of Biology and Center for Molecular Biology of RNA, University of California, Santa Cruz, Santa Cruz, California 95064, USA

Analyzing vertebrate genomes requires rapid mRNA/DNA and cross-species protein alignments. A new tool, BLAT, is more accurate and 500 times faster than popular existing tools for mRNA/DNA alignments and 50 times faster for protein alignments at sensitivity settings typically used when comparing vertebrate sequences. BLAT's speed stems from an index of all nonoverlapping K-mers in the genome. This index fits inside the

BLAT = BLAST-like Alignment Tool

- Rapid searches by INDEXING the entire genome
- Works best with high similarity matches
- See documentation and publication for details
 - Kent, WJ. *Genome Res.* 2002. 12:656 and “Help”

BLAT Tool Interface

■ Make choices

Human BLAT Search

BLAT Search Genome

Genome:	Assembly:	Query type:	Sort output:	Output type:
>uc002gij.3 (TP53)	length=2591			

```
gatgggattgggggtttccctcccatgtgctcaagactggcgctaaaagttttgagctt
c
g
g
t
c
g
g
g
c
g
g
c
g
t
t
g
t
t
t
g
c
c
a
a
c
t
g
g
c
c
a
a
g
a
c
t
g
c
c
c
t
g
t
g
c
a
g
c
t
g
t
g
g
g
t
t
g
a
t
t
c
c
a
c
c
c
c
c
g
c
c
```

submit submit I'm feeling lucky clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence: Browse... submit file

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.

For locating PCR primers, use [In-Silico PCR](#) for best results instead of BLAT.

- Paste one or more sequences
- FASTA for more than one

*DNA limit 25000 bases
Protein limit 10000 aa
25 total sequences*

■ Or upload

Short seqs

BLAT Results with Hyperlinks

Human BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	uc002gij.3	2581	1	2591	2591	100.0%	17	-	7571720	7590868	19149
browser details	uc002gij.3	177	2158	2436	2591	83.1%	1	+	45290354	45290634	281
browser details	uc002gij.3	176	2134	2433	2591	85.6%	10	-	27408468	27408791	324
browser details	uc002gij.3	174	2141	2437	2591	83.7%	2	+	27384674	27384975	302
browser details	uc002gij.3	174	2134	2436	2591	87.6%	10	+	67312526	67312836	311
browser details	uc002gij.3	173	2148	2431	2591	87.4%	10	+	71133346	71133631	286
browser details	uc002gij.3	173	2149	2504	2591	84.0%	10	+	65420577	65421000	424
browser details	uc002gij.3	172	2153	2433	2591	83.4%	3	+	27600067	27600347	281
browser details	uc002gij.3	165	2160	2444	2591	88.4%	X	-	122127686	122127972	287
browser details	uc002gij.3	160	2152	2435	2591	83.2%	2	-	109493652	109493934	283
browser details	uc002gij.3	160	2137	2434	2591	84.0%	1	-	225930110	225930396	287
browser details	uc002gij.3	160	2144	2437	2591	83.5%	10	+	15559328	15559614	287
browser details	uc002gij.3	160	2138	2552	2591	82.9%	9	-	131379044	131379531	488
browser details	uc002gij.3	160	2158	2435	2591	82.2%	4	-	139925816	139926096	281
browser details	uc002gij.3	160	2134	2414	2591	84.3%	10	-	12095247	12095528	282
browser details	uc002gij.3	160	2127	2434	2591	86.0%	2	+	170700494	170700797	304
browser details	uc002gij.3	159	2128	2435	2591	85.4%	13	-	103001995	103002383	308
browser details	uc002gij.3	26	2128	2154	2591	100.0%	3	-	27607611	27607638	28
browser details	uc002gij.3	26	2408	2437	2591	93.4%	X	+	47169213	47169242	30
browser details	uc002gij.3	26	2273	2304	2591	90.7%	5	+	7460469	7460500	32
browser details	uc002gij.3	25	2358	2389	2591	82.8%	2	+	124842060	124842089	30
browser details	uc002gij.3	23	2353	2379	2591	92.6%	X	-	100332288	100332314	27
browser details	uc002gij.3	23	2323	2345	2591	100.0%	X	+	47169722	47169744	23
browser details	uc002gij.3	22	2369	2404	2591	80.6%	20	-	33243008	33243043	36
browser details	uc002gij.3	21	2182	2202	2591	100.0%	2	+	38998603	38998623	21
browser details	uc002gij.3	21	2347	2367	2591	100.0%	1	+	199938363	199938383	21

Annotations: A blue arrow on the left points to the 'browser' and 'details' links with the text 'go to browser/viewer'. Another blue arrow points to the 'QUERY' column with the text 'go to alignment detail'. A blue box labeled 'sorting' is placed over the 'SCORE' column.

- Results with demo sequences, settings default; sort = Query, Score
 - Score is a count of matches—higher number, better match
- Click [browser](#) to go to Genome Browser image location (next slide)
- Click [details](#) to see the alignment to genomic sequence (2nd slide)

BLAT Results, Alignment Details

Alignment of uc002gij.3

[uc002gij.3](#)

[Human chr17](#)

[block1](#)

[block2](#)

[block3](#)

[block4](#)

[block5](#)

[block6](#)

[block7](#)

[block8](#)

[block9](#)

[block10](#)

[block11](#)

[together](#)

Alignment of uc002gij.3 and chr17:7571720-7590868

Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence (often splice sites).

cDNA uc002gij.3

Your query

```
GATGGGATTG GGGTTTTCCC CTCCCATGTG CTCGAAGACTG GCGCTAAAAG 50
TTTTGAGCTT CTCAAAAGTC TAGAGCCACC GTCCAGGGAG CAGGTAGCTG 100
CTGGGCTCCG GGGACACTTT GCGTTCGGGC TGGGAGCGTG CTTTCCACGA 150
CGGTGACACG CTTCCCTGGA TTTCCTGGA TTGgtaagc tcttgactga
CCATGGAGGA GCCGCAGT GAAACATTTT CAGACCTT CCCCTTGCAG TCCCAAGG
CCCTTGCAG TCCCAAGG TTGAACAATG GTTCACTG CCAGAGGCTG
CTCCCTCG GGCCTTGC AAGAAACCTA CCAGGGCA GGGACAGCCA
AGTCTGTG GTTTTGCCAA CTGGCCAC CACCCCGGCC CGGCACCC
CAGCACATGA CCGGAGCT AGATAGCGAT GGTCTGGG ATTTGCGTGT
GGAGTATT GTGGTGCCCT ATGAGCCG CTACAACACT
ATGTGTAT CCATCTCTAC CATCATCG CGGAACAGCT
TTGAGGTC CACAGAGGAA GAGAATCT CCCCAGGGAG
CACTAAGC CAGCCAAAAG AGAAACCC TGGCGGTGAG
CGCTTCG TCAAGGATGC CCAGGCTG ACCCACCTGA
AGTCCAAATCATGTTCAAG ACAGAAGG
```

Genomic chr17 (reverse strand):

```
tccccaaactc catttccctt gcttctctcg gcaggcggat tacttgcct 7590919
tacttgtcat ggcgactgtc cagctttgtg ccaggagcct cgcaggggtt 7590869
GATGGGATTG GGGTTTTCCC CTCCCATGTG CTCGAAGACTG GCGCTAAAAG 7590819
TTTTGAGCTT CTCAAAAGTC TAGAGCCACC GTCCAGGGAG CAGGTAGCTG 7590769
CTGGGCTCCG GGGACACTTT GCGTTCGGGC TGGGAGCGTG CTTTCCACGA 7590719
tctctcttga gtcacgggct ctcggctcog tgtattttca
atcgctggg c
a
g
ttgaacgct a
aaagcgct t
gaaataaaag a
acattgagaa tcatagctg tatattttag agcccatggc
aaactggggc tccattccga aatgatcatt tgggggtgat
caagctgcta aggtcccaca acttccggac ctttgtcctt
tctttccagg cagcccccg ctcgctaga tggagaaaat
gctgtcagtc gtggaagtga gaagtgctaa accagggggt
gccgaggagg accgtcgcaa tctgagaggc ccggcagccc
tggctccaca ttacatttc tgctctt
tttgccggag cagctcacta ttcaccog
gaaaatgtcc ttaggcog tctctt
ttctccgctt gcatttctt ttctggat
aggcagggtt atttgtttg atgcaaaact caatccctcc
atggtgtgcc ccaccccg ggtcgcctgc aacctaggc
```

Genomic region, with color cues

yours genomic

Side by Side Alignment

Side by Side Alignment

```
0000001 gatgggattggggttttccctccatgtgctcaagactggcgctaaaaa 0000050
<<<<<< ||| <<<<<<<<
7590868 gatgggattggggttttccctccatgtgctcaagactggcgctaaaaa 7590819

0000051 ttttgagcttctcaaaagtctagagccaccgtccaggagcaggtagctg 0000100
<<<<<< ||| <<<<<<<
7590818 ttttgagcttctcaaaagtctagagccaccgtccaggagcaggtagctg 7590769

0000101 ctgggctcggggacactttgcgttcgggctgggagcgtgctttccaca 0000150
<<<<<< ||| <<<<<<<
7590768 ctgggctcggggacactttgcgttcgggctgggagcgtgctttccaca 7590719

0000151 cggtgacacgcttccctggattgg 0000174
<<<<<< ||| <<<<<<<
7590718 cggtgacacgcttccctggattgg 7590695

0000224 cagccagactgcctccgggtcactgccatggaggagccgcagtcagatc 0000274
<<<<<< || <<<<<<<
7579891 cagccagactgcctccgggtcactgccatggaggagccgcagtcagatc 7579841

0000275 ct 0000276
<<<<<< || <<<<<<<
7579840 ct 7579839

0000277 acttctctgaaacacagctctg 0000298
```

Cruzar informações de tracks de
anotação e selecionar regiões de DNA
de interesse (Table Browser)

The Table Browser

The screenshot shows the UCSC Genome Bioinformatics website. The main navigation bar includes links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. The 'Tools' menu is open, listing various tools such as Blat, Table Browser, Variant Annotation Integrator, Data Integrator, Gene Sorter, Genome Graphs, In-Silico PCR, LiftOver, VisiGene, and Other Utilities. The 'Table Browser' option is highlighted with a mouse cursor. The left sidebar contains a list of tools including Genome Browser, Blat, Table Browser, Gene Sorter, In Silico PCR, Genome Graphs, Galaxy, VisiGene, Utilities, Downloads, Release Log, Custom Tracks, Cancer Browser, Microbial Genomes, ENCODE, and Neandertal. The main content area features an 'About the UCSC Genome Browser' section with introductory text and a 'DONATE NOW' button. Below this is a 'News' section with a Twitter and Facebook icon, and a 'News Archives' link. A news item dated 29 June 2015 is visible, titled 'GENCODE Genes Now the Default Gene Set on the Human (GRCh38/hg38) Assembly'. A large URL 'http://genome.ucsc.edu' is overlaid at the bottom of the page.

Tipos de polimorfismos frequentemente encontrados nos genomas

- Variações em uma única base:
Single Nucleotide Polymorphisms, SNPs
- Variações que afetam múltiplas bases no genoma:
Ex. Diferenças no número de trechos repetitivos no DNA

SNP short tandem repeat (STR)



Indivíduo 1 GTACTAGACTACTACTACTACTACTCTGGTG...
5 repeats

Indivíduo 2 GTACAAGACTACTACTACTACTACTACTCTGGTG...
6 repeats

Indivíduo 3 GTACAAGACTACTACTACTACTACTACTACTACTCTGGTG...
7 repeats

Discovery using the Table Browser

Table 1 | Features of trinucleotide expansion in humans

Disease	Sequence	Location	Parent of origin of expansion	Repeat number (normal)	Repeat number (pre-mutation)	Repeat number (disease)	Somatic instability
Diseases with coding TNRs							
DRPLA	CAG	ATN1 (exon 5)	P	6–35	35–48	49–88	Yes
HD	CAG	HTT (exon 1)	P	6–29	29–37	38–180	Yes
OPMD	GCN	PABPN1 (exon 1)	P and M	10	12–17	>11	None found in tissue tested (hypothalamus)
SCA1	CAG	ATXN1 (exon 8)	P	6–39	40	41–83	Yes
SCA2	CAG	ATXN2 (exon 1)	P	<31	31–32	32–200	Unknown
SCA3 (Machado–Joseph disease)	CAG	ATXN3 (exon 8)	P	12–40	41–85	52–86	Unknown
SCA6	CAG	CACNA1A (exon 47)	P	<18	19	20–33	None found
SCA7	CAG	ATXN7 (exon 3)	P	4–17	28–33	>36 to >460	Yes
SCA17	CAG	TBP (exon 3)	P > M	25–42	43–48	45–66	Yes
SMBA	CAG	AR (exon 1)	P	13–31	32–39	40	None found
Diseases with non-coding TNRs							
DM1	CTG	DMPK (3' UTR)	M	5–37	37–50	<50	Yes
DM2	CCTG	CNBP (intron 1)	Uncertain	<30	31–74	75–11,000	Yes
FRAX-E	GCC	AFF2 (5' UTR)	M	4–39	40–200	>200	Unknown
FRDA	GAA	FXN (intron 1)	Recessive	5–30	31–100	70–1,000	Yes
FXS	CGG	FMR1 (5' UTR)	M	6–50	55–200	200–4,000	Yes
HDL2	CTG	JPH3 (exon 2A)	M	6–27	29–35	36–57	Unknown
SCA8	CTG	ATXN8OS (3' UTR)	M	15–34	34–89	89–250	Unknown
SCA10	ATTCT	ATXN10 (intron 9)	M and P (smaller changes with M)	10–29	29–400	400–4,500	Yes
SCA12	CAG	PPP2R2B (5' UTR)	M and P (more unstable with P)	7–28	28–66	66–78	None found

Are there other trinucleotide repeats of the sequence “CAG” that might be of interest to extend the study?

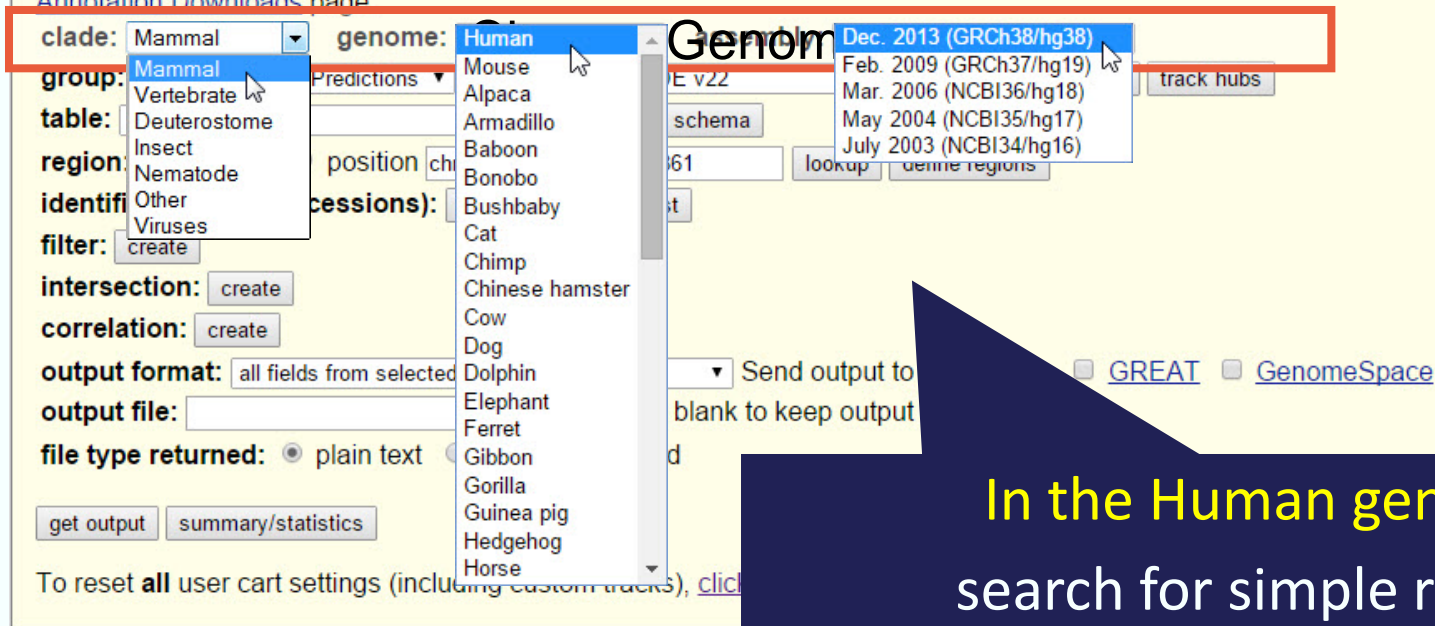
The Table Browser

The screenshot shows the UCSC Genome Bioinformatics website. The main navigation bar includes links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. The 'Tools' menu is open, listing various utilities such as Blat, Table Browser, Variant Annotation Integrator, Data Integrator, Gene Sorter, Genome Graphs, In-Silico PCR, LiftOver, VisiGene, and Other Utilities. The 'Table Browser' option is highlighted with a blue bar and a mouse cursor. A dark blue arrow points from the 'Table Browser' text to the highlighted menu item. Another dark blue arrow points from the 'Table Browser' text to the 'Table Browser' link in the left-hand navigation sidebar. The main content area features an 'About the UCSC Genome Browser' section with introductory text and a 'DONATE NOW' button. Below this is a 'News' section with a Twitter and Facebook icon, and a 'News Archives' link. A news item dated 29 June 2015 is visible, titled 'GENCODE Genes Now the Default Gene Set on the Human (GRCh38/hg38) Assembly'. A dark blue box at the bottom of the page contains the URL 'http://genome.ucsc.edu'.

Table Browser: Choose Genome

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.



In the Human genome,
search for simple repeats
in the entire genome
with the exact sequence "CAG"
and get table data.

Table Browser: Choose Table to Search

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal genome: Human assembly: Dec. 2013 (GRCh38/hg38)

group: Mapping and Sequencing track: RepeatMasker
Genes and Gene Models RepeatMasker Viz.
table: Phenotype and Literature

region: mRNA and EST
Expression
identifi Regulation
filter: Comparative Genomics
Variation
interse Repeats
correla All Tracks
All Tables

output format: all fields from selected table Send output to GREAT GenomeSpace
output file: (leave blank to keep output
file type returned: plain text gzip compressed

get output summary/statistics

To reset all user cart settings (including custom tracks), [click here](#)

Choose Data Table

Custom Tracks + Hubs

In the Human genome, search for simple repeats in the entire genome with the exact sequence "CAG" and get table data.

Table Browser: Describe Table

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: genome: assembly:
 group: track:

table:

Describe table

region: genome position

identifiers (names/accessions):

filter:

intersection:

correlation:

output format:

output file: (leave blank to use default)

file type returned: plain text gzip compressed

To reset all user cart settings (including custom tracks), [click here](#)

Schema for Simple Repeats - Simple Tandem Repeats by TRF

Database: hg38 Primary Table: simpleRepeat Row Count: 1,014,212 Data last updated: 2013-12-29
 Format description: Describes the Simple Tandem Repeats

field	example	SQL type	description
bin	585	smallint(5) unsigned	Indexing field to speed chromosome range queries.
chrom	chr1	varchar(255)	Reference sequence chromosome or scaffold
chromStart	10000	int(10) unsigned	Start position in chromosome
chromEnd	10468	int(10) unsigned	End position in chromosome
name	trf	varchar(255)	Simple Repeats tag name
period	6	int(10) unsigned	Length of repeat unit
copyNum	77.2	float	Mean number of copies of repeat
consensusSize	6	int(10) unsigned	Length of consensus sequence
perMatch	95	int(10) unsigned	Percentage Match
perIndel	3	int(10) unsigned	Percentage Indel
score	789	int(10) unsigned	Alignment Score = 2*match-7*mismatch-7*indel; minscore=50
A	33	int(10) unsigned	Percent of A's in repeat unit
C	51	int(10) unsigned	Percent of C's in repeat unit
G	0	int(10) unsigned	Percent of G's in repeat unit
T	15	int(10) unsigned	Percent of T's in repeat unit
entropy	1.43	float	Entropy
sequence	TAACCC	longBlob	Sequence of repeat unit element

Sample Rows

bin	chrom	chromStart	chromEnd	name	period	copyNum	consensusSize	perMatch	perIndel	score	A	C	G	T	entropy	sequence
585	chr1	10000	10468	trf	6	77.2	6	95	3	789	33	51	0	15	1.43	TAACCC
585	chr1	10627	10800	trf	29	6	29	100	0	346	13	38	47	0	1.43	AGGCGCGCCGCGCGC
585	chr1	10757	10997	trf	76	3.2	76	95	2	434	17	30	45	6	1.73	GGCGCAGGCGCAG
585	chr1	11225	11447	trf	117	1.9	121	80	14	273	12	32	33	20	1.9	CGCCCCCTGCTGGC
585	chr1	11271	11448	trf	61	2.9	61	82	4	187	12	32	34	20	1.9	AGTGGTGGCACGCC
585	chr1	11283	11448	trf	62	2.7	61	82	2	199	12	33	33	20	1.9	CGCCCCCTGCTGGC

Table Browser: Choose Region to Search

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: genome: assembly:
group: track:
table:

region: genome position coordinates intervals from regions

Define regions

Choose Region to Search

Enter region definition

Paste regions: Or upload file: No file selected.

Region definitions are ordinary 3- or 4-field bed file format. For example:

```
# comment lines can be included starting with the # symbol
chrX 151073054 151173000
chrX 151183000 151190000 optionalRegionName
chrX 151283000 151290000
chrX 151383000 151390000
```

There is a limit of 1,000 defined regions. Using the upload file function will replace any existing region definitions. For notation purposes only, it is not used in the table browser. These are 0-relative coordinates, first base on a chromosome is 1.

Coordinates can also be entered in the form:

```
chrX:151,283,001-151,290,000 optionalRegionName
```

These are 1-relative coordinates, first base on a chromosome is 1.

GREAT GenomeSpace

In the Human genome,
search for simple repeats
in the entire genome
with the exact sequence "CAG"
and get table data.

Table Browser: Quick Summary of Expected Hits

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: **genome:** **assembly:**

group: **track:**

table:

region: genome position

identifiers (names/accessions):

filter:

intersection:

correlation:

output format:

output file: (leave blank to keep output)

file type returned: plain text gzip compressed

Data Summary

To reset **all** user cart settings (including custom tracks), [click here](#).

Simple Repeats (simpleRepeat) Summary Statistics

item count	1,014,212
item bases	146,785,521 (4.81%)
item total	335,386,791 (11.00%)
smallest item	25
average item	144
biggest item	500
smallest score	
average score	
biggest score	80

Region and Timing Statistics

region	genom
bases in region	3,209
bases in gaps	159
load time	
calculation time	
free memory time	
filter	
intersection	

~ 1 million
Simple Repeats
annotated
in the Human genome

Table Browser: Filter to Refine Search

Table Browser

Use this program to retrieve the data associated with a track in the browser. For help in using this application see [Using the Table Browser](#). For more information and sample queries, and the OpenHelix Table Browser. For more complex queries, you may want to use [Galaxy](#) or our [public](#) annotation enrichments, send the data to [GREAT](#). Send data to [GREAT](#) for the list of contributors and usage restrictions associated with the [Annotation Downloads](#) page.

clade: genome: assembly:
group: track:
table:
region: genome position
identifiers (names/aliases):
filter: **Create Filter**
intersection:
correlation:
output format:
output file: (leave blank to keep default)
file type returned: plain text gzip compressed

To reset all user cart settings (including custom tracks), [click here](#).

Filter on Fields from hg38.simpleRepeat

bin	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	
chrom	does	match	<input type="text" value="*"/>	AND
chromStart	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
chromEnd	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
name	does	match	<input type="text" value="*"/>	
period	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	
copyNum	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	
consensusSize	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	
perMatch	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	
perIndel	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	
score	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
A	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
C	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
G	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
T	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
entropy	is	<input type="text" value="ignored"/>	<input type="text" value="0"/>	AND
sequence	does	match	<input type="text" value="CAG"/>	

Submit Filter

Delete wildcard "*" type in CAG (uppercase)

with the exact sequence "CAG" and get table data.

Table Browser: Output Data

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: genome: assembly:
group: track:

table:

region: genome position

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: GREAT GenomeSpace

output file: (leave blank to keep output file)

file type returned: plain text gzip compressed

To reset **all** user cart settings (including custom tracks), [click here](#)

External tools

Output data

In the Human genome,
search for simple repeats
in the entire genome
with the exact sequence "CAG"
and get table data.

Table Browser: Output Formats

Table Browser

Use this program to retrieve the data associated with a track. For help in using this application, see the [Table Browser](#) page. For general information and sample queries, and the OpenHelix Wiki. For more complex queries, you may want to use [Galaxy](#). For annotation enrichments, send the data to [GREAT](#). See the [Table Browser](#) page for the list of contributors and usage restrictions associated with the [Annotation Downloads](#) page.

clade: **genome:**
group: **track:**
table:
region: genome position
identifiers (names/accessions):
filter:
intersection:
correlation:

output format: Send output to [Galaxy](#) [GREAT](#) [GenomeSpace](#)
output file: (click to keep output in browser)
file type return:

[hyperlinks to Genome Browser](#)
 To reset all user cart settings (including custom tracks), [click here](#).

#filter: simpleRepeat.sequence = 'CAG'

#bin	chrom	chromStart	chromEnd	name	period	copyNum	consensusSize	perMatch	perIndel
848	chr1	34593457	34593520	trf	3	21	85	6	83
1136	chr1	72282938	72282981	trf	3	14.3	85	0	59
1206	chr1	81501781	81501833	trf	3	17.3	87	0	77
1479	chr1	117210212	117210244	trf	3	10.7	100	0	64
1735	chr1	150770967	150771006	trf	3	13	83	0	51
2173	chr1	208267465	208267493	trf	3	9.3	100	0	56
708	chr2	16239536	16239584	trf	3	16	100	0	96
740	chr2	20425653	20425682	trf	3	9.7	100	0	58
1261	chr2	88627211	88627211	trf	3	8.3	100	0	50
1984	chr2	183404468	183404468	trf	3	27.3	84	0	101
2361	chr2	232847490	232847490	trf	3	15.3	78	17	51
858	chr3	35871111	35871111	trf	3	8.7	100	0	52
1459	chr3	114610054	114610116	trf	3	20.7	79	0	70
1559	chr3	127774708	127774734	trf	3	8.7	100	0	52
2023	chr3	188491547	188491596	trf	3	16.3	78	0	53
608	chr4	3074876	3074940	trf	3	96	119	34	34
1186	chr4	78870928	78871011	trf	3	27.7	90	0	130
1239	chr4	85749945	85749977	trf	3	10.7	93	0	55
1448	chr4	113116686	113116724	trf	3	13	83	5	51
1837	chr4	164161545	164161575	trf	3	10	100	0	60
924	chr5	44467844	44467882	trf	3	13	91	5	60
1435	chr5	111414381	111414413	trf	3	10.7	93	0	55
931	chr6	45422681	45422747	trf	3	22	90	0	105

Text Fields

Output formats

Table Browser: File Format Outputs

Table Browser

Use this program to ...
 covered by a track.
 general information a
 For more complex qu
 annotation enrichme
 for the list of contribu
[Annotation Download](#)

clade:

group:

table:

region: genome

identifiers (names/

filter:

intersection:

correlation:

output format: Send output to Galaxy GREAT

output file: (click to keep output in browser)

file type return

- sequence
- GTF - gene transfer format
- BED - browser extensible data
- custom track
- hyperlinks to Genome Browser

To reset all user cart settings (including custom tracks), [click here](#).

chr	hg38_simpleRepeat	exon	34593458	34593520	83.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr1	hg38_simpleRepeat	exon	72282939	72282981	59.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup1";
chr1	hg38_simpleRepeat	exon	81501782	81501833	77.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup2";
chr1	hg38_simpleRepeat	exon	117210213	117210244	64.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup3";
chr1	hg38_simpleRepeat	exon	150770968	150771006	51.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup4";
chr1	hg38_simpleRepeat	exon	208267466	208267493	56.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup5";
chr2	hg38_simpleRepeat	exon	16239537	16239584	96.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr2	hg38_simpleRepeat	exon	20425654	20425682	58.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup1";
chr2	hg38_simpleRepeat	exon		36	50.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup2";
chr2	hg38_simpleRepeat	exon		550	101.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup3";
chr2	hg38_simpleRepeat	exon		536	51.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup4";
chr2	hg38_simpleRepeat	exon		37	52.000000	.	.	gene_id	"trf"; transcript_id	"trf_dup4";
chr3	hg38_simpleRepeat	exon	114610055	114610116	70.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr3	hg38_simpleRepeat	exon	127774709	127774734	52.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr3	hg38_simpleRepeat	exon	188491548	188491596	53.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr4	hg38_simpleRepeat	exon	3074877	3074940	119.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr4	hg38_simpleRepeat	exon	78870929	78871011	130.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr4	hg38_simpleRepeat	exon	85749946	85749977	55.000000	.	.	gene_id	"trf"; transcript_id	"trf";
chr4	hg38_simpleRepeat	exon	113116687	113116724	51.000000	.	.	gene_id	"trf"; transcript_id	"trf";

chr	34593457	34593520	trf	83
chr1	72282938	72282981	trf	59
chr1	81501781	81501833	trf	77
chr1	117210212	117210244	trf	64
chr1	150770967	150771006	trf	51
chr1	208267465	208267493	trf	56
chr2	16239536	16239584	trf	96
chr2	20425653	20425682	trf	58
chr2			trf	50
chr2			trf	101
chr2			trf	51
chr3			trf	52
chr3	114610054	114610116	trf	70
chr3	127774708	127774734	trf	52
chr3	188491547	188491596	trf	53
chr4	3074876	3074940	trf	119
chr4	78870928	78871011	trf	130
chr4	85749945	85749977	trf	55
chr4	113116686	113116724	trf	51
chr4	164161545	164161575	trf	60
chr5	44467844	44467882	trf	60
chr5	111414381	111414413	trf	55
chr6	45422681	45422747	trf	105
chr6	123895239	123895270	trf	53
chr6	156778268	156778321	trf	88
chr6	160218912	160218977	trf	58
chr7	39339688	39339719	trf	62
chr7	114631524	114631556	trf	55
chr7	149471291	149471324	trf	66
chr8	9190624	9190658	trf	59
chr8	100810508	100810540	trf	64
chr8	133055824	133055872	trf	96
chr9	389592	389621	trf	58
chr9	113425354	113425386	trf	55

Table Browser: Custom Track Output

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal genome: Human

group: Repeats track: Simple Repeat

table: simpleRepeat describe table schema

region: genome position chr9:133252000-13328086

identifiers (names/accessions): paste list upload list

filter: edit clear

intersection: create

correlation: create

output format: all fields from selected table

output file: all fields from selected table

file type returned: selected fields from primary and related table
sequence
GTF - gene transfer format
BED - browser extensible data
custom track
hyperlinks to Genome Browser

get output summary

To reset all user cart settings (including custom track...

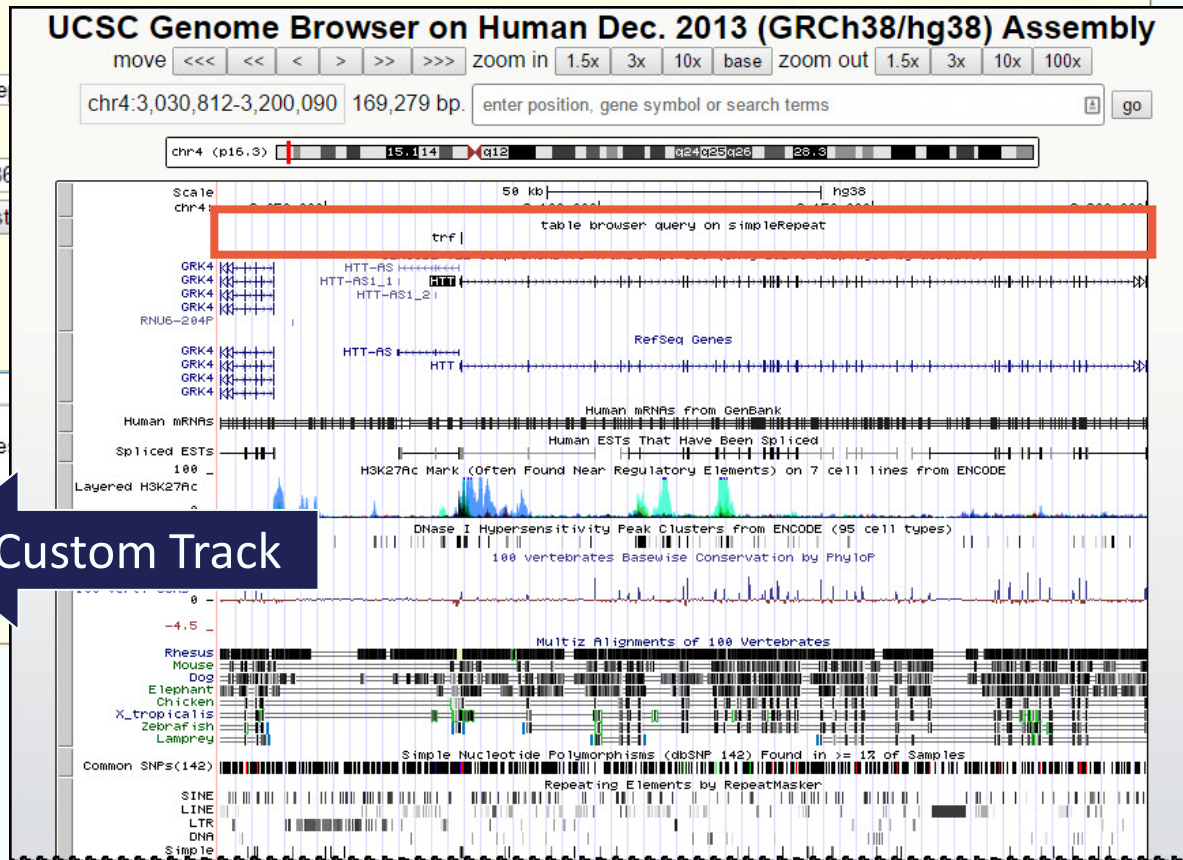


Table Browser: Hyperlinks Output

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersection covered by a track. For help in using this application see [Using the Table Browser](#) for a description of general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine gene annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse tools for the list of contributors and usage restrictions associated with these data. All tables can be downloaded from the [Annotation Downloads](#) page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Repeats **track:** Simple Repeats

table: simpleRepeat

region: genome position chr9:133252000-133280861

identifiers (names/accessions):

filter:

intersection:

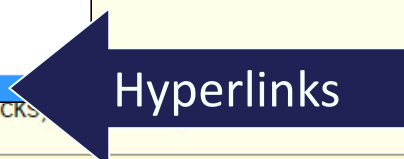
correlation:

output format: all fields from selected table Galaxy GREAT (click to keep output in browser)

output file: all fields from selected table

file type returned: selected fields from primary and related tables
sequence
GTF - gene transfer format
BED - browser extensible data
custom track
[hyperlinks to Genome Browser](#)

To reset all user can settings (including custom tracks, ...)



Hyperlinks

Hyperlinks to Genome Browser

[trf at chr1:34593458-34593520](#)
[trf at chr1:72282939-72282981](#)
[trf at chr1:81501782-81501833](#)
[trf at chr1:117210213-117210244](#)
[trf at chr1:150770968-150771006](#)
[trf at chr1:208267466-208267493](#)
[trf at chr2:16239537-16239584](#)
[trf at chr2:20425654-20425682](#)
[trf at chr2:88627212-88627236](#)
[trf at chr2:183404469-183404550](#)
[trf at chr2:232847491-232847536](#)
[trf at chr3:35871112-35871137](#)
[trf at chr3:114610055-114610116](#)
[trf at chr3:127774709-127774734](#)
[trf at chr3:188491548-188491596](#)
[trf at chr4:3074877-3074940](#)
[trf at chr4:78870929-78871011](#)
[trf at chr4:85749946-85749977](#)
[trf at chr4:113116687-113116724](#)
[trf at chr4:164161546-164161575](#)
[trf at chr5:44467845-44467882](#)
[trf at chr5:111414382-111414413](#)
[trf at chr6:45422682-45422747](#)
[trf at chr6:123895240-123895270](#)
[trf at chr6:156778269-156778321](#)
[trf at chr6:160218913-160218977](#)
[trf at chr7:39339689-39339719](#)
[trf at chr7:114631525-114631556](#)
[trf at chr7:119471292-119471324](#)

Table Browser: Obtaining Output

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the [Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for

Browser tutorial for a narrated presentation of the software features and usage

82 simple repeats in the genome match our filter criteria, out of ~1 million

Simple Repeats (simpleRepeat) Summary Statistics

item count	82
item bases	3,304 (0.00%)
item total	3,304 (0.00%)
smallest item	25
average item	40
biggest item	93
smallest score	50
average score	67
biggest score	130

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: all fields from selected table

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

Data Summary

To reset all user cart settings (including custom tracks), [click here](#).

Adding name creates file on desktop, leaving blank creates output in browser. (exception: custom track)

bases in gaps	159,950,299
load time	3.64
calculation time	0.41
free memory time	0.00
filter	on
intersection	off

Table Browser: Output Configuration

Table Browser

Use this program to retrieve the data associated with a track. For help in using this application, see the [Table Browser](#) page. For more information and sample queries, and for more complex queries, you may want to see the [Table Browser](#) page. For the list of contributors and usage restrictions, see the [Table Browser](#) page.

filter: simpleRepeat.sequence = 'CAG'

clade: **genome:**

group: **track:**

table:

region: genome position

identifiers (names/accessions):

filter:

intersection:

correlation:

output format:

output file:

file type returned:

To reset all user interface settings (including custom tracks), [click here](#).

#bin	chrom	chromStart	chromEnd	name	period	copyNum	consensusSize	perMatch	perIndel	score	A	C	G	T	entropy	sequence
848	chr1	34593457	34593520	trf	3	21	3	85	6	83	34	26	38	0	1.57	CAG
1136	chr1	72282938	72282981	trf	3	14.3	3	85	0	59	39	32	27	0	1.57	CAG
1206	chr1	81501781	81501833	trf	3	17.3	3	87	0	77	38	32	28	0	1.57	CAG
1479	chr1	117210212	117210244	trf	3	10.7	3	100	0	64	34	34	31	0	1.58	CAG
1735	chr1	150770967	150771006	trf	3	13	3	83	0	51	33	30	35	0	1.58	CAG
2173	chr1	208267465	208267493	trf	3	9.3	3	100	0	56	32	35	32	0	1.58	CAG
708	chr2	16239536	16239584	trf	3	16	3	100	0	96	33	33	33	0	1.58	CAG
740	chr2	20425653	20425682	trf	3	9.7	3	100	0	58	34	34	31	0	1.58	CAG
1261	chr2	88627211	88627236	trf	3	8.3	3	100	0	50	32	36	32	0	1.58	CAG
1984	chr2	183404468	183404550	trf	3	27.3	3	84	0	101	28	30	37	3	1.74	CAG
2361	chr2	232847490	232847536	trf	3	15.3	3	78	17	51	28	36	30	4	1.76	CAG
858	chr3	35871111	35871137	trf	3	8.7	3	100	0	52	34	34	30	0	1.58	CAG
1459	chr3	114610054	114610116	trf	3	20.7	3	79	0	70	40	30	27	1	1.66	CAG
1559	chr3	127774708	127774734	trf	3	8.7	3	100	0	52	34	34	30	0	1.58	CAG
2023	chr3	188491547	188491596	trf	3	16.3	3	78	0	53	38	28	28	4	1.75	CAG
608	chr4	3074876	3074940	trf	3	21.3	3	96	0	119	34	34	31	0	1.58	CAG
1186	chr4	78870928	78871011	trf	3	27.7	3	90	0	130	38	33	27	0	1.57	CAG
1239	chr4	85749945	85749977	trf	3	10.7	3	93	0	55	31	34	34	0	1.58	CAG
1448	chr4	113116686	113116724	trf	3	13	3	83	5	51	31	31	34	2	1.72	CAG
1837	chr4	164161545	164161575	trf	3	10	3	100	0	60	33	33	33	0	1.58	CAG
924	chr5	44467844	44467882	trf	3	13	3	91	5	60	34	34	28	2	1.71	CAG
1435	chr5	111414381	111414413	trf	3	10.7	3	93	0	55	37	34	28	0	1.57	CAG
931	chr6	45422681	45422747	trf	3	22	3	90	0	105	37	33	28	0	1.58	CAG
1530	chr6	123895239	123895270	trf	3	10.3	3	92	0	53	35	35	29	0	1.58	CAG
1781	chr6	156778268	156778321	trf	3	17.7	3	92	0	88	37	33	28	0	1.58	CAG
1807	chr6	160218912	160218977	trf	3	21.7	3	82	0	58	40	26	27	6	1.8	CAG
885	chr7	39339688	39339719	trf	3	10.3	3	100	0	62	32	35	32	0	1.58	CAG
1459	chr7	114631524	114631556	trf	3	10.7	3	93	0	55	37	34	28	0	1.57	CAG
1725	chr7	149471291	149471324	trf	3	11	3	100	0	66	33	33	33	0	1.58	CAG
655	chr8	9190624	9190658	trf	3	93	0	59	29	35	35	0	1.58	CAG		
1354	chr8	100810508	100810540	trf	3	10.7	3	100	0	64	34	34	31	0	1.58	CAG
1600	chr8	133055824	133055872	trf	3	16	3	100	0	96	33	33	33	0	1.58	CAG
587	chr9	389592	389621	trf	3	9.7	3	100	0	58	34	34	31	0	1.58	CAG
1450	chr9	113425354	113425386	trf	3	10.7	3	93	0	55	31	34	34	0	1.58	CAG
734	chr9	19585690	19585721	trf	3	10.3	3	100	0	62	32	35	32	0	1.58	CAG

- Table Browser Introduction
- Table Browser: Set up and Filters
- **Table Browser: Intersections**
- Custom Tracks: Table Browser & User Generated
- Summary
- Exercises

UCSC Table Browser: <http://genome.ucsc.edu>

Table Browser: Intersecting Data

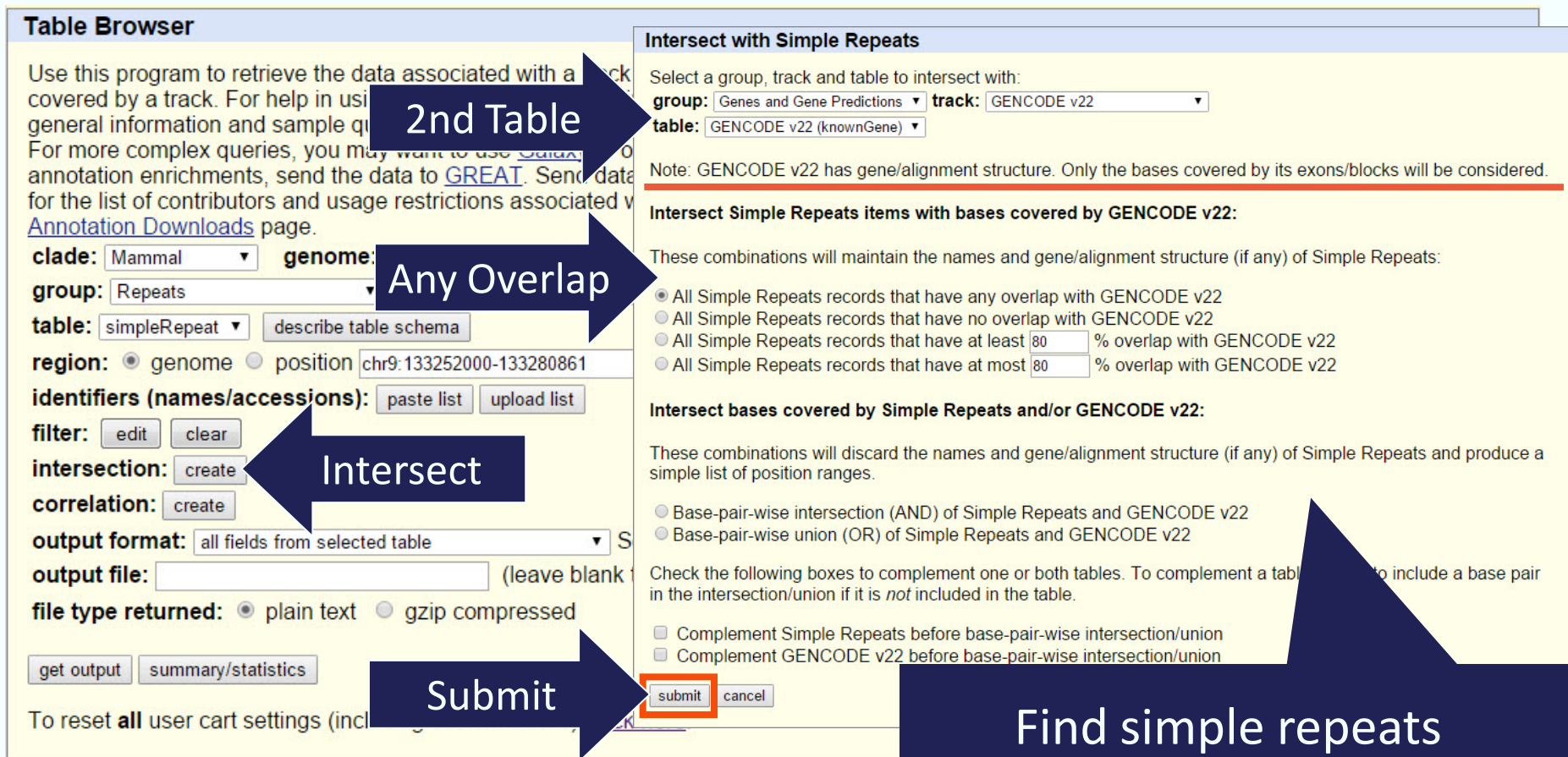


Table Browser

Use this program to retrieve the data associated with a track covered by a track. For help in using the program, see the general information and sample queries. For more complex queries, you may want to use [GREAT](#). Send data for the list of contributors and usage restrictions associated with [Annotation Downloads](#) page.

clade: genome:

group: table:

region: genome position

identifiers (names/accessions):

filter:

intersection: **Intersect**

correlation:

output format:

output file: (leave blank)

file type returned: plain text gzip compressed

To reset all user cart settings (including the list of tracks)

Intersect with Simple Repeats

Select a group, track and table to intersect with:
group: track:
table:

Note: GENCODE v22 has gene/alignment structure. Only the bases covered by its exons/blocks will be considered.

Intersect Simple Repeats items with bases covered by GENCODE v22:

These combinations will maintain the names and gene/alignment structure (if any) of Simple Repeats:

- All Simple Repeats records that have any overlap with GENCODE v22
- All Simple Repeats records that have no overlap with GENCODE v22
- All Simple Repeats records that have at least % overlap with GENCODE v22
- All Simple Repeats records that have at most % overlap with GENCODE v22

Intersect bases covered by Simple Repeats and/or GENCODE v22:

These combinations will discard the names and gene/alignment structure (if any) of Simple Repeats and produce a simple list of position ranges.

- Base-pair-wise intersection (AND) of Simple Repeats and GENCODE v22
- Base-pair-wise union (OR) of Simple Repeats and GENCODE v22

Check the following boxes to complement one or both tables. To complement a table, you must include a base pair in the intersection/union if it is *not* included in the table.

- Complement Simple Repeats before base-pair-wise intersection/union
- Complement GENCODE v22 before base-pair-wise intersection/union

2nd Table

Any Overlap

Submit

Find simple repeats (sequence = CAG) within known genes and get hyperlinks.

Table Browser: Intersecting Data Narrows Search

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for the list of contributors and usage restrictions associated with these data. See the [Annotation Downloads](#) page.

clade: genome: assembly:
group: track:
table:
region: genome position
identifiers (names/accessions):
filter:

intersection with knownGene:

correlation:

output format: Send output to [Galaxy](#)

output file: (leave blank to keep output)

file type returned: plain text gzip compressed

Note: The all fields and selected tracks are not available when an intersection has been performed.

To reset all user cart settings (including custom tracks), [click here](#).

Simple Repeats (simpleRepeat) Summary Statistics

item count	30
item bases	1,349 (0.00%)
item total	1,349 (0.00%)
smallest item	5
average item	
biggest item	
smallest score	
average score	
biggest score	

Summary

Filtered simple repeats,
intersected (overlapping)
with known genes

Table Browser: Get Hyperlinks

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections, or to generate a BED file. For help in using this application see [Using the Table Browser](#) for a general overview and sample queries, and the OpenHelix Table Browser [tutorial](#) for a more detailed overview. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To download the data for annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with the list of contributors and usage restrictions associated with these data. All tables can be downloaded from the [Annotation Downloads](#) page.

clade: genome: assembly:

group: track: [manage custom tracks](#)

table: [describe table schema](#)

region: genome position [lookup](#) [define regions](#)

identifiers (names/accessions): [paste list](#) [upload list](#)

filter: [edit](#) [clear](#)

intersection with knownGene: [edit](#) [clear](#)

correlation: [create](#)

output format: Send output to [Galaxy](#) [GREAT](#) [Table Browser](#)

output file: (leave blank to keep output in browser)

file type returned: [BED - browser extensible data](#) [GTF - gene transfer format](#) [sequence](#) [custom track](#)

Note: The all tracks are sorted by genomic position. When an intersection is found, the output will include a [hyperlink to Genome Browser](#) for each region.

[get output](#)

get output

hyperlinks

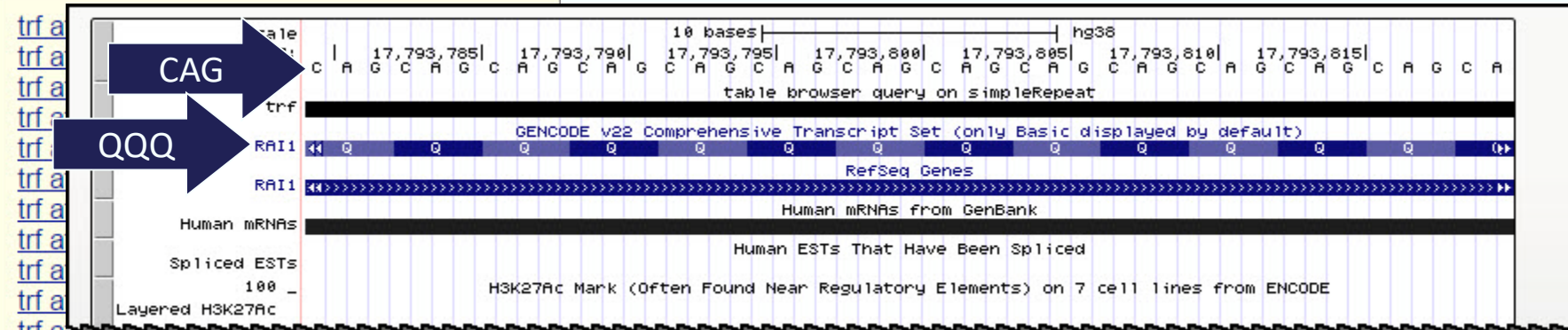
To reset all user interface settings (including custom tracks), [click here](#).

Hyperlinks to Genome Browser

[trf at chr2:88627212-88627236](#)
[trf at chr2:232847491-232847536](#)
[trf at chr4:3074877-3074940](#)
[trf at chr4:78870929-78871011](#)
[trf at chr6:45422682-45422747](#)
[trf at chr6:156778269-156778321](#)
[trf at chr7:39339689-39339719](#)
[trf at chr7:114631525-114631556](#)
[trf at chr9:113425355-113425386](#)
[trf at chrX:107603188-107603268](#)
[trf at chrX:150471047-150471078](#)
[trf at chr11:6641515-6641540](#)
[trf at chr11:9091445-9091485](#)
[trf at chr11:65557855-65557894](#)
[trf at chr11:130428218-130428247](#)
[trf at chr12:6936717-6936775](#)
[trf at chr12:132606198-132606231](#)
[trf at chr15:43618669-43618701](#)
[trf at chr15:74543949-74543988](#)
[trf at chr15:99712505-99712538](#)
[trf at chr16:67195891-67195947](#)
[trf at chr16:67879864-67879912](#)
[trf at chr16:69693627-69693667](#)
[trf at chr17:4887671-4887701](#)
[trf at chr17:7024701-7024730](#)
[trf at chr17:17136248-17136282](#)
[trf at chr17:17793780-17793820](#)
[trf at chr19:45770205-45770266](#)
[trf at chr19:49423274-49423301](#)
[trf at chr22:20566470-20566562](#)

Table Browser: Exploring the results

Hyperlinks to Genome Browser



Human Gene RAI1 (uc002grm.4) Description and Page Index

Description: Homo sapiens retinoic acid induced 1 (RAI1), mRNA. (from RefSeq NM_030665)

RefSeq Summary (NM_030665): This gene is located within the Smith-Magenis syndrome region on chromosome 17. It is highly similar to its mouse counterpart and is expressed at high levels mainly in neuronal tissues. The protein encoded by this gene includes a polymorphic polyglutamine tract in the N-terminal domain. Expression of the mouse counterpart in neurons is induced by retinoic acid. This gene is associated with both the severity of the phenotype and the response to medication in schizophrenic patients. [provided by RefSeq, Jul 2008]. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##Evidence-Data-

START## Transcript exon combination :: AY172136.1 [ECO:0000332] RNAseq introns :: single sample supports all introns SAMEA1966682, SAMEA1968189 [ECO:0000348] ##Evidence-Data-END##

Gencode Transcript: ENST00000353383.4

Transcript (Including UTRs)

Position: chr17:17,681,473-17,811,453 **Size:** 129,981 **Total Exon Count:** 6 **Strand:** +

Coding Region:

RAI1, chr17: 17793780...

Tutorial - UCSC Genome Browser

- 1) executar uma análise semelhante a descrita nos slides anteriores para identificar genes candidatos a doenças associadas a expansão de trinucleotídeos usando a ferramenta “Table Browser”
- 2) Anotação de novos RNAs não codificadores usando informações genômicas no UCSC Genome Browser
- 3) Integração com dados de bulk e single cell RNAseq e dados clínicos

Instruções para confecção do relatório:

Responder as perguntas de cada tutorial. Podem usar nas respostas as imagens geradas na análise.

O relatório deve ser entregue através do e-disciplinas.

Tutorial - Banco de dados

1. Anotação de novos RNAs não codificadores usando informações genômicas no UCSC Genome Browser

Em um estudo que reconstruiu o transcrito de tumores pancreáticos (Paixão et al., 2022) foram identificados cinco RNAs não codificadores longos (lncRNAs) super-expressos que ainda não estão descritos no catálogo de referência de genes humanos (GENCODE)

A sequência dos lncRNAs (em formato fasta) estão disponíveis na página da disciplina (lncRNAs_upPDAC.fasta)

- 1) utilize o programa BLAT para alinhar as sequências na versão mais atual do genoma humano (dica: escolha o melhor alinhamento, maior score, para prosseguir)
- 2) reporte as coordenadas de mapeamento dos transcritos incluindo: cromossomo, coordenadas de início e fim, orientação da fita do DNA.
- 3) reporte o número de exons e o tamanho de cada transcrito maduro
- 4) reporte o nome e a distância (em bases) do gene mais próximo anotado na vizinhança dos RNAs.
- 5) Ative tracks com informações de elementos regulatórios da expressão gênica (ex. promotores, enhancers, ilhas CpG) e busque evidências que corroborem a transcrição do lncRNA candidato. Reporte se encontrou elementos regulatórios na proximidade do início de transcrição de algum dos RNAs.
- 6) Ative tracks com informações de conservação evolutiva e busque evidências que corroborem a existência de seleção da sequência dos lncRNAs candidatos. Reporte se observou trechos conservados na sequência de algum dos RNAs.

2. Integração com dados de bulk e single cell RNAseq e dados clínicos

Na mesma análise foram identificados vinte lncRNAs já anotados no GENCODE, alguns dos quais com expressão alterada já descrita em tumores de pâncreas ou outros contextos tumorais. Para avançar no estudo desses RNAs é importante saber em que tipo celular são expressos (vários tipos celulares no microambiente tumoral). Bancos de dados de expressão gênica que integram dados de “bulk tissue” (RNAseq) e células únicas (scRNAseq) possibilitam acesso a dados experimentais já anotados.

Os identificadores dos RNAs (gene symbol) estão na tabela abaixo.

LINC01559	AC005550.3	RP11-350J20.12	MIR210HG
LINC02577	CCAT1	UNC5B-AS1	LINC01614
RP11-460N11.2	RP11-284F21.10	UCA1	RP11-38M8.1
LINC00675	RP3-340N1.2	MMP12	ACER3
LINC01133	RP11-284F21.9	RP11-395B7.4	LINC00941

1. No portal Cancer Single-cell Expression Map <https://ngdc.cncb.ac.cn/cancerscem/> busque por informações sobre a expressão desses lncRNAs em estudos de single cell e bulk tissue RNA seq.
2. Para quantos desses RNAs foi possível encontrar informações?
3. Algum desses RNAs tem um padrão de expressão que sugira ser preferencialmente expresso em tumores de pâncreas? Nota: os projetos com amostras de pâncreas tem o nome de PDAC ou TCGA-PAAD
4. No portal Gepia2 <http://gepia2.cancer-pku.cn/> verifique se a expressão de algum dos RNAs acima está correlacionada com a sobrevida global (overall survival) ou sobrevida livre da doença.